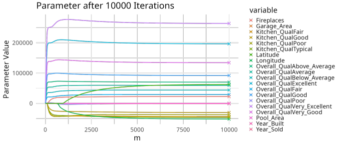
Introduction to Machine Learning

Gradient Boosting: CWB and GLMs



Learning goals

- Understand relationship of CWB and GLM

RELATION TO GLM

In the simplest case we use linear models (without intercept) on single features as base learners:

$$b_j(x_j, \theta) = \theta x_j \quad \text{for } j = 1, 2, \dots, p \quad \text{and with } b_j \in \mathcal{B}_j = \{\theta x_j \mid \theta \in \mathbb{R}\}.$$

This definition will result in an ordinary **linear regression** model.

- In the limit, boosting algorithm will converge to the maximum likelihood solution.
- By specifying loss as NLL of exponential family distribution with an appropriate link function, CWB is equivalent to (regularized) **GLM**.

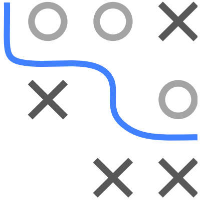
RELATION TO GLM

But: We do not *require* an exponential family distribution and we can - in principle - apply it to any differentiable loss

Usually we do not let the boosting model converge fully, but use **early stopping** for the sake of regularization and feature selection.

Even though resulting model looks like a GLM, we do not have valid standard errors for our coefficients, so cannot provide confidence or prediction intervals or perform tests etc. $\rightarrow$ post-selection inference.

EXAMPLE: LOGISTIC REGRESSION WITH CWB



Fitting a logistic regression (GLM with a Bernoulli distributed response) requires the specification of the loss as function as

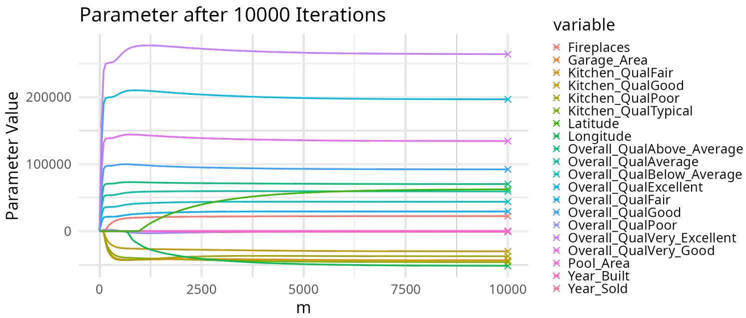
$$L(y, f(\mathbf{x})) = -y \cdot f(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))), \; y \in \{0, 1\}$$

Note that CWB (as gradient boosting in general) predicts a score $f(\mathbf{x}) \in \mathbb{R}$. Squashing the score $f(\mathbf{x})$ to $\pi(\mathbf{x}) = s(f(\mathbf{x})) \in [0, 1]$ corresponds to transforming the linear predictor of a GLM to the response domain with a link function s:

- $s(f(\mathbf{x})) = (1 + \exp(-f(\mathbf{x})))^{-1}$ for logistic regression.
- $s(f(\mathbf{x})) = \Phi(f(\mathbf{x}))$ for probit regression with Φ the CDF of the standard normal distribution.

EXAMPLE: CWB PARAMETER CONVERGENCE

The following figure shows the parameter values for $m \leq 10000$ iterations as well as the estimates from a linear model as crosses (GLM with normally distributed errors):



Throughout the fitting of CWB, the parameters estimated converge to the GLM solution. The used data set is Ames Housing.